CLAIMS

1. A method of back-off modelling for use in named entity recognition of a text, comprising, for an initial pattern entry from the text:

relaxing one or more constraints of the initial pattern entry;

determining if the pattern entry after constraint relaxation has a valid form; and

moving iteratively up the semantic hierarchy of the constraint if the pattern entry after constraint relaxation is determined not to have a valid form.

2. A method according to claim 1, wherein moving iteratively up the semantic hierarchy of the constraint if the pattern entry after constraint relaxation is determined not to have a valid form comprises:

moving up the semantic hierarchy of the constraint;

relaxing the constraint further; and

returning to determining if the pattern entry after constraint relaxation has a valid form.

3. A method according to claim 1 or 2, further comprising:

determining if a constraint in the pattern entry, after relaxation, also has a valid form; and

moving iteratively up the semantic hierarchy of the constraint if the constraint in the pattern entry after constraint relaxation is determined not to have a valid form.

4. A method according to claim 3, wherein moving iteratively up the semantic hierarchy of the constraint if the constraint in the pattern entry after constraint relaxation is determined not to have a valid form comprises:

moving up the semantic hierarchy of the constraint;

relaxing the constraint further; and

returning to determining if a constraint in the pattern entry after constraint relaxation has a valid form.

5. A method according to any one of the preceding claims, wherein if a constraint is relaxed, the constraint is dropped entirely from the pattern entry if the relaxation reaches the root of the semantic hierarchy.

6. A method according to any one of the preceding claims, further comprising terminating if a near optimal frequently occurring pattern entry is reached to replace the initial pattern entry.

7. A method according to any one of the preceding claims, further comprising selecting the initial pattern entry for back-off modelling if it is not a frequently occurring pattern entry in a lexicon.

8. A method of inducing patterns in a pattern lexicon comprising a plurality of initial pattern entries with associated occurrence frequencies, the method comprising:

identifying one or more initial pattern entries in the lexicon with lower occurrence frequencies; and

relaxing one or more constraints of individual ones of the identified one or more initial pattern entries to broaden the coverage of the identified one or more initial pattern entries.

9. A method according to claim 8, further comprising creating the pattern lexicon of initial pattern entries from a training corpus.

10. A method according to claim 8 or 9, further comprising merging individual ones of the constraint relaxed initial pattern entries with similar pattern entries in the lexicon to form a more compact pattern lexicon.

11. A method according to claim 9 or 10, wherein the entries in the compact pattern lexicon are generalised as much as possible within a given similarity threshold.

12. A method according to any one of claims 8 to 11, further comprising:

determining if the pattern entry after constraint relaxation has a valid form; and

moving iteratively up the semantic hierarchy of the constraint if the pattern entry after constraint relaxation is determined not to have a valid form.

13. A method according to claim 12, wherein moving iteratively up the semantic hierarchy of the constraint if the pattern entry after constraint relaxation is determined not to have a valid form comprises:

moving up the semantic hierarchy of the constraint;

relaxing the constraint further; and

returning to determining if the pattern entry after constraint relaxation has a valid form.

14. A method according to claim 12 or 13, further comprising:

determining if a constraint in the pattern entry, after relaxation, also has a valid form; and

moving iteratively up the semantic hierarchy of the constraint if the constraint in the pattern entry after constraint relaxation is determined not to have a valid form.

15. A method according to claim 14, wherein moving iteratively up the semantic hierarchy of the constraint if the constraint in the pattern entry after constraint relaxation is determined not to have a valid form comprises:

moving up the semantic hierarchy of the constraint;

relaxing the constraint further; and

returning to determining if a constraint in the pattern entry after constraint relaxation has a valid form.

16. A decoding process in a rich feature space comprising a method according to any one of claims 1 to 7.

17. A training process in a rich feature space comprising a method according to any one of claims 8 to 15.

18. A system for recognising and classifying named entities within a text, comprising:

feature extraction means for extracting various features from a document;

recognition kernel means to recognise and classify named entities using a Hidden Markov Model; and

back-off modelling means for back-off modelling by constraint relaxation to deal with data sparseness in a rich feature space.

19. A system according to claim 18, wherein the back-off modelling means is operable to provide a method of back-off modelling according to any one of claims 1 to 7.

20. A system according to claim 18 or 19, further comprising a pattern induction means for inducing frequently occurring patterns.

21. A system according to claim 20, wherein the pattern induction means is operable to provide a method of inducing patterns according to any one of claims 8 to 15.

22. A system according to anyone of claims 18 to 21, wherein said various features are extracted from words within the text and the discourse of the text, and comprise one or more of:

a. deterministic features of words, including capitalisation or digitalisation;
b. semantic features of trigger words;
c. gazetteer features, which determine whether and how the current word string appears in a gazetteer list;
d. discourse features, which deal with the phenomena of name alias; and
e. the words themselves.

23. A feature set for use in back-off modelling in a Hidden Markov Model, during named entity recognition, wherein the feature sets are arranged hierarchically to allow for data sparseness.